2

NPS55-89-01

NAVAL POSTGRADUATE SCHOOL

Monterey, California

INFERRING FINITE-TIME PERFORMANCE IN THE M/G/1 QUEUEING MODEL

P. A. JACOBS
D. P. GAVER

JANUARY 1989

Approved for public release; distribution is unlimited.

Prepared for:
Office of Naval Research
Arlington, VA 22217

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral R. C. Austin
Superintendent

H. Shull
Provost

This report was prepared in conjunction with research conducted for the Office of Naval Research and funded by the Naval Postgraduate School.

Reproduction of all or part of this report is authorized.

This report was prepared by:

Patricia A. Jacobs
PATRICIA A. JACOBS
Professor of Operations Research

Donald P. Gaver
DONALD P. GAVER
Professor of Operations Research

Reviewed by:

P. Purdue
PETER PURDUE
Professor and Chairman
Department of Operations Research

Released by:

K. T. Marshall
KNEALE T. MARSHALL
Dean of Information and
Policy Sciences

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3 DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution is unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
NPS55-89-01

5 MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Naval Postgraduate School

6b OFFICE SYMBOL *(If applicable)*
Code 55

7a. NAME OF MONITORING ORGANIZATION
Office of Naval Research

6c. ADDRESS *(City, State, and ZIP Code)*
Monterey, CA 93943-5000

7b. ADDRESS *(City, State, and ZIP Code)*
Arlington, VA 22217

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
Naval Postgraduate School

8b OFFICE SYMBOL *(If applicable)*

9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
O&MN, Direct Funding

8c. ADDRESS *(City, State, and ZIP Code)*
Monterey, CA 93943-5000

10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO.

11. TITLE *(Include Security Classification)*
INFERRING FINITE-TIME PERFORMANCE IN THE M/G/1 QUEUEING MODEL

12. PERSONAL AUTHOR(S)
Jacobs, P. A. and Gaver, D. P.

13a. TYPE OF REPORT
Technical

13b TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Year, Month, Day)*
1989 January

15 PAGE COUNT
29

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Empirical Transform; Laplace transform of the virtual waiting of the M/G/1 queue; Exponential Approximation; Brownian motion with drift. (JES)

19 ABSTRACT *(Continue on reverse if necessary and identify by block number)*

A single server is approached by a stream of Poisson arrivals with known arrival rate. The service times are assumed to be independent identically distributed with unknown distribution. One has available a finite sample of service times obtained by observing the system. A nonparametric approach is taken towards estimating the expected waiting time encountered by a new arriving customer at a finite time t, $E[W_t]$ both for stable and unstable systems. The estimator uses approximations to $E[W_t]$ and an empirical version of the well known Laplace transform of $E[W_t]$ for the M/G/1 queue.

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21 ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
Donald P. Gaver

22b TELEPHONE *(Include Area Code)*
(408)646-2605

22c OFFICE SYMBOL
Code 55Gv

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

☆ U.S. Government Printing Office 1986—606-243

INFERRING FINITE-TIME PERFORMANCE IN THE M/G/1 QUEUEING MODEL

P.A. Jacobs
D.P. Gaver

1. INTRODUCTION AND SUMMARY

The M/G/1 model for a single server approached by a Poisson (λ) stream of arrivals with independent but otherwise arbitrarily distributed service times is a standard tool in operations research. It has been used to model situations occurring in road traffic, computer and communication performance evaluation and the military, and has been modified to accommodate priority and server breakdown situations, as well as both heavy and light traffic situations ($\rho = \lambda E[S]$ either near, but below, unity, and near, but above, zero).

Although the M/G/1 model is rather simplistic, little beyond the moments of its long-run or steady-state distribution are easily available in anything like closed (and simple) analytical form readily evaluated numerically. In the case of the exponential service time M/M/1 system. the complete analytical solution for transition probabilisties in terms of series of Bessel functions has been long known, and numerical transform inversion by Middleton (1979) has provided some useful tables describing the manner in which steady state is approached. Recent work of Abate and Whitt (1989,1988) provides some handy approximations, particularly in the M/M/1 case. The work of Asmussen and Thorisson (1988) and Newell (1971,

1982) furnishes diffusion approximations and heavy traffic results as well.

In this paper we explore the following operationally relevant

Problem: A single server is approached by a stream of arrivals to be modelled as Poisson. One has available a sample of service times obtained by observing the system. Using the arrival rate (assumed known, at least initially) and the observed service times, estimate the expected waiting time encountered by a new arriving customer at finite time t. Furthermore, provide an assessment of estimation uncertainty, e.g. confidence limits.

The above is just one prototypical question to be asked--it is the one addressed in this paper. Other questions could concern the probability that an arrival at t will wait for at least time w, or that the customer waiting time has never exceeded w in time t (starting from an empty system). There are many other such measures of system performance that are relevant. We concentrate here on inferring the mean waiting time at finite time t, and do so non-parametrically, i.e. without specifying, and estimating parameters in a small conventional parametric family such as the Gamma. Our approach is to work with the well-known Laplace transform of the mean (virtual) waiting time for the M/G/1 system: cf. Takács (1962) and Gaver and Jacobs (1987), an empirical version of which can be obtained by utilizing the empirical transform of the busy period, itself depending upon the empirical transform of the service time df, derived from the observed data.

We look upon our present results as exploratory and tentative, but of an accuracy useful until refinements are available.

A recent paper by Tengels (1988) provides mathematically rigorous asymptotic (large sample size) treatment of certain aspects of our nonparametric inference problem. Application of bootstrap methods, cf. Efron (1979) can yield further insights into the inference at the price of computer intensivity.

2. MATHEMATICAL FORMULATION

Consider an M/G/1 queue with known arrival rate λ and independent identically distributed service times, S, having nth moment $E[S^n] < \infty$. Suppose there are no customers either waiting or being served at time 0.

Let W_t be the virtual waiting time at time t and put (conditional on $W_0 = 0$)

$$\psi_W(s) = \int_0^\infty e^{-st} \cdot E[W_t] \, dt, \quad s \geq 0 . \tag{1}$$

We know that

$$\psi_W(s) = \frac{\rho - 1}{s^2} + p_{00}(s) \frac{1}{s} ; \tag{2}$$

where:

$$p_{00}(s) = \frac{1}{s + \lambda[1 - b(s)]} , \tag{3}$$

b(s) being the Laplace transform of a busy period duration; b(s) is the smallest positive solution of

$$b(s) = \hat{F}_S(s + \lambda(1 - b(s))) ; \tag{4}$$

$\hat{F}_S$ is the Laplace-Stieltjes transform of the service time. See Takács (1962) and Gaver and Jacobs (1986).

Two approximations to $E[W_t | W_0 = 0]$ will be detailed below; one for the case of the traffic intensity $\lambda E[S] < 1$; the other for the case of the traffic intensity $\lambda E[S] \geq 1$. Both are based on the transforms (2), (3), and (4).

3. APPROXIMATIONS FOR STABLE QUEUES: EXPONENTIAL APPROACH

In this subsection, it is assumed that $\rho = \lambda E[S] < 1$. In this case there is a limiting expected virtual waiting time

$$\bar{W}_\infty = \lim_{t \to \infty} E[W_t] = \frac{\lambda E[S^2]}{2(1 - \rho)} , \tag{5}$$

with

$$\rho = \lambda E[S] , \tag{6}$$

the traffic intensity.

Our choice of the form of the approximation is based on an observation of Odoni and Roth (1983), and on relaxation time calculations made by Keilson (1979), Newell (1971) and others. The explicit approximation of the exponential:

$$E[W_t] \approx \bar{W}_\infty (1 - e^{-\beta t})$$

or

$$\frac{\bar{W}_\infty - E[W_t]}{\bar{W}_\infty} \approx e^{-\beta t} . \tag{7}$$

It should be noted that in many cases the exponential approach to the limiting value is not exact, being only an approximation, cf. Asmussen and Thorissan (1988). Setting

$$\int_0^\infty se^{-st} \frac{E[W_t] - \bar{W}_\infty}{\bar{W}_\infty} dt = - \int_0^\infty se^{-st} e^{-\beta t} dt \tag{8}$$

we obtain

$$\frac{s\psi_W(s) - \bar{W}_\infty}{\bar{W}_\infty} = - \frac{s}{s + \beta} \tag{9}$$

which is now solved for β at selected s-values:

$$1 + \beta(\tfrac{1}{s}) = - \frac{\bar{W}_\infty}{s\nu_W(s) - \bar{W}_\infty} ;$$

so

$$\beta(\tfrac{1}{s}) = \frac{-\bar{W}_\infty - [s\psi_W(s) - \bar{W}_\infty]}{s\psi_W(s) - \bar{W}_\infty} . \qquad (10)$$

Finally

$$\beta = \frac{-s^2\psi_W(s)}{s\psi_W(s) - \bar{W}_\infty} . \qquad (11)$$

The approximation to $E[W_t]$ is

$$E[W_t] \approx \bar{W}_\infty(1 - e^{-\beta t}) , \qquad (12)$$

where β is evaluated as above at $s = \frac{1}{t}$. It turns out that the β-values so obtained are not constant but are (extremely) slowly-changing functions of t, as should be the case from asymptotic analysis.

Table 1 reports the values of the approximating $E[W_t]$ for an M/M/1 queue with $\lambda = 1$ and $E[S] = .95$. The approximating values are compared to values obtained by Middleton (1979), who used a numerical inversion of the Laplace transform ψ_W due to Stehfest (1970). The approximation is very good for the times less than or equal to 100. For the times greater than 100 the approximation is within 11% of the true, and conservatively slightly overestimate of the true mean.

Table 1

Exponential Approximation
M/M/1 Queue: $\lambda = 1$, $E[S] = .95$

Time t	True $E[W_t]$	Exponential Approximation
20	3.9	3.7
40	5.5	5.4
60	6.6	6.6
80	7.4	7.6
100	8.1	8.4
400	13.2	14.6
800	15.5	17.0
1200	16.6	17.7

Note: $\bar{W}_\infty = 18.05$

Table 2 reports the values of the approximating $E[W_t]$ for an M/G/1 queue with $\lambda = 1$ and gamma service times with shape parameter $\alpha = .2$ and scale parameter $\beta = 4.5$. Once again the approximation is very good. The values of the true $E[W_t]$ are from Middleton (1979).

Table 2

Exponential Approximation
M/G/1 Queue: $\lambda = 1$
Service Time Distribution Gamma shape parameter
$\alpha = .2$, scale parameter $\beta = 4.5$, $E[S] = .90$

Time	True	Exponential
t	$E[W_t]$	Approximation
4.9	2.3	2.2
9.7	3.7	3.4
29.2	6.9	6.7
48.6	8.8	8.8
97.2	11.9	12.5
194.4	15.3	16.9
486	20.1	21.9
972	22.6	23.7
1458	23.5	23.8
1944	23.9	24.1

Note: $\bar{W}_{\infty} = 24.3$

The numerical results obtained suggest that in general the approximation is biased slightly on the high side for large t. It is thus conservative in the sense that predictions of expected waiting times tend to be slightly overstated.

4. THE HEAVY TRAFFIC EXPONENTIAL APPROXIMATION (HTE)

The most computationally-intensive part of the exponential approximation is the numerical evaluation of the transform of the busy period b(s) using equation (4). This computation has been done by search. In this section another approximation is described which uses a heavy traffic approximation to obtain b(s), and is computationally less arduous.

Let $\{B(t)\}$ be a Brownian motion with drift $\nu = \lambda E[S] - 1$ and infinitesimal variance $\sigma^2 = \lambda E[S^2]$ where S is a generic service time. Let T_x be the first passage time to state 0 given $B(0) = x$. Put

$$\phi(s;x) = E[e^{-sT_x}] \quad . \tag{13}$$

It is known that

$$\phi(s:x) = e^{\alpha(s)x} \tag{14}$$

where

$$\alpha(s) = \frac{-\nu - \left|\sqrt{\nu^2 + 2\sigma^2 s}\right|}{\sigma^2} \tag{15}$$

Of course this function is explicit and readily evaluated, and no search is needed.

Approximating the virtual waiting time process by the Brownian motion with drift, the proposed approximation for the transform of the busy period is

$$b_{BM}(s) = E[e^{\alpha(s)S}] \quad , \tag{16}$$

the Laplace-Stieljes transform of the service time distribution evaluated at $\alpha(s)$. The remainder of the approximaticn is the same as the exponential approximation. Table 3 shows the values of the heavy traffic approximation for the M/G/1 queue with $\lambda = 1$ and gamma service time distribution with shape parameter $\alpha = 0.2$ (very long-tailed) and scale parameter $\beta = 4.5$. The heavy traffic approximation is not as good as the exponential one and tends to overestimate $E[W_t]$. However, the approximation is still practically adequate, and is far easier to compute than is the "exact" exponential approximation of Section 2.

Table 3

Unsaturated Heavy Traffic Approximation
M/G/1 Queue: $\lambda = 1$, Gamma Service Times
Shape parameter $\alpha = 0.2$, scale parameter $\beta = 4.5$, $E[S] = .90$

Time	True	Heavy Traffic
t	$E[W_t]$	Approximation
4.9	2.3	2.5
9.7	3.7	3.9
29.2	6.9	7.5
48.6	8.8	9.8
97.2	11.9	13.6
194.4	15.3	17.9
972	22.6	24.1
1944	23.9	24.3

Note: $\bar{W}_\infty = 24.3$

5. THE BUSY PERIOD APPROXIMATION

This approach is based on approximating the rate of approach of the probability o an empty system, $P_{00}(t)$, to its eventual value. Since the process of alternating busy and idel periods can be expressed in terms of an alternating renewal process, methods of Gaver and Jacobs (1988) can be applied.

Note:

$$s \cdot \psi_w(s) = \frac{\rho - 1}{s} + p_{00}(s) \ , \tag{17}$$

where

$$p_{00}(s) = \frac{1}{s} \frac{1}{1 + \lambda[\frac{1 - b(s)}{s}]} \tag{18}$$

and

$$p_{00}(s) = \int_0^{\infty} e^{-st} P\{X(t) = 0 | X(0) = 0\} dt \tag{19}$$

where X(t) is the number of customers waiting or being served at time t. Assume the queue is stable; then

$$\lim_{t \to \infty} P\{X(t) = 0 | X(0) = 0\} = \frac{E[Idle]}{E[Idle] + E[Busy]} = 1 - \rho \ . \tag{20}$$

where Idle is the length of an idle period and Busy is the length of a busy period. Approximate as follows focussing on the probability that the system is empty, approximates as follows in terms of the parameter β_b:

$$P\{X(t) = 0 | X(0) = 0\} \approx (1 - \rho) + \rho e^{-\beta_b t} \ ; \tag{21}$$

from this comes

$$\begin{aligned} E[W_t] &\approx (\rho - 1)t + (1 - \rho)t + \frac{\rho}{\beta_b}(1 - e^{-\beta_b t}) \\ &= \frac{\rho}{\beta_b}(1 - e^{-\beta_b t}) \ . \end{aligned} \tag{22}$$

To find β_b:

$$P_{00}(s) = \frac{(1 - \rho)}{s} + \frac{\rho}{\beta_b + s} \tag{23}$$

$$P_{00}(s) = \frac{1}{s}\,\frac{1}{1 + \lambda[\frac{1 - b(s)}{s}]} = \frac{1}{s}\,\frac{1}{1 + \frac{\rho}{1 - \rho}C(s)} \tag{24}$$

where

$$C(s) = \frac{[1 - b(s)]}{sE[B]} \quad \text{and} \quad \lambda E[B] = \frac{\rho}{1 - \rho} \tag{25}$$

Solving for β_b results in

$$\beta_b = \frac{sC(s)[\frac{\rho}{1 - \rho} + 1]}{1 - C(s)} = \frac{sC(s)[\frac{1}{1 - \rho}]}{1 - C(s)}\ , \tag{26}$$

and evaluate at $s = \frac{1}{t}$.

Table 4 shows values for the busy period approximation for an M/M/1 queue with $\lambda = 1$ and $\rho = .95$. The approximate values are below the true values for most times t. This approximation is not as accurate as is the exponential approximation of Section 2, no longer being conservative.

Table 4
Busy Period Approximation
M/M/1 Queue
$\lambda = 1$, E[S] = .95

Time t	True $E[W_t]$	Approximate $E[W_t]$
20	3.9	4.1
40	5.5	5.4
60	6.6	6.3
80	7.4	7.0
100	8.1	7.5
400	13.2	11.6
800	15.5	13.5
1200	16.6	14.6

6. THE UNSTABLE OR SATURATED QUEUE EXPONENTIAL APPROXIMATION

Rewriting, we have

$$\psi_W(s) = \frac{\rho - 1}{s^2} + \frac{1}{s^2} \frac{1}{1 + \lambda[\frac{1 - b(s)}{s}]} \tag{27}$$

$$= \frac{\rho - 1}{s^2} + \frac{1}{s} \frac{1}{s + \lambda[1 - b(s)]} \ . \tag{28}$$

In this subsection we assume that $\rho > 1$. When this is true the distribution of the length of the busy period is not honest, and $b(0) < 1$. We will write $b(0)\ b^{\#}(s) = b(0)\ [\frac{b(s)}{b(0)}]$, where $b^{\#}(s)$ is now the transform of an honest random variable. Now

$$\psi_W(s) = \frac{\rho - 1}{s^2} + \frac{1}{s} \frac{1}{s + \lambda[1 - b(0)\ b^{\#}(s)]} \ . \tag{29}$$

For small s

$$\psi_W(s) \approx \frac{\rho - 1}{s^2} + \frac{1}{s} \frac{1}{\lambda[1 - b(0)]} \ . \tag{30}$$

Thus for large t

$$E[W_t] \approx (\rho - 1)t + \frac{1}{\lambda[1 - b(0)]} \ . \tag{31}$$

Further

$$\psi_W(s) = \frac{\rho - 1}{s^2} + \frac{1}{s}\phi(s) \tag{32}$$

where

$$\phi(s) = \frac{1}{s + \lambda[1 - b(0)b^{\#}(s)]} = \frac{1}{s + \lambda - \lambda b(0)b^{\#}(s)} \ .$$

Multiplying through by $s + \lambda$ yields

$$(s + \lambda)\phi(s) = \frac{1}{1 - \frac{\lambda}{s + \lambda} b(0)b^{\#}(s)} \ ;$$

or

$$(s + \lambda)\phi(s) = 1 + (s + \lambda)\phi(s)\left[\frac{\lambda}{s + \lambda} b(0) b^{\#}(s)\right] \; ; \tag{33}$$

and

$$\phi(s) = \frac{1}{s + \lambda} + \frac{\lambda}{s + \lambda} b(0) b^{\#}(s)\phi(s) \; . \tag{34}$$

Letting $\phi(s) = \int_0^\infty e^{-st} f(t)dt$

$$f(t) = e^{-\lambda t} + \int_0^\infty f(t-u)h(u)du$$

where

$$\gamma(s) \equiv \int_0^\infty e^{-st} h(t)dt = b(0) \frac{\lambda}{s + \lambda} b^{\#}(s) \tag{35}$$

and

$$h(0) = b(0) < 1 \; .$$

Thus f satisfies a renewal-type equation with a degenerate distribution and it can be shown (cf. Feller (1966))

$$f(t) \approx ce^{-\beta t} \quad \text{as} \quad t \to \infty \; . \tag{36}$$

Thus, since

$$\psi_W(s) = \frac{\rho - 1}{s^2} + \frac{1}{s}\phi(s)$$

$$E[W_t] \approx (\rho - 1)t + \frac{c}{\beta}(1 - e^{-\beta t}) \tag{37}$$

as $t \to \infty$. Comparing (36) with (31) indicates that

$$\frac{c}{\beta} = \frac{1}{\lambda[1 - b(0)]} \; . \tag{38}$$

In the saturated case the exponential approximation is computed as follows:

$$E[W_t] - (\rho - 1)t \approx a\,(1 - e^{-\beta t})$$

where

$$a = \frac{1}{\lambda[1 - b(0)]} .$$

Thus

$$\frac{E[W_t] - (\rho - 1)t - a}{a} = -e^{-\beta t} . \tag{39}$$

Taking Laplace transforms

$$\frac{s\upsilon_W(s) - (\frac{\rho - 1}{s}) - a}{a} = - \frac{1}{1 + \beta(\frac{1}{s})} \tag{40}$$

which leads to

$$\beta = \frac{-s^2\upsilon_W(s) + (\rho - 1)}{s\upsilon_W(s) - (\frac{\rho - 1}{s}) - 1} \tag{41}$$

since

$$a = \frac{1}{\lambda[1 - b(0)]} . \tag{42}$$

7. THE SATURATED QUEUE HEAVY TRAFFIC EXPONENTIAL APPROXIMATION

Once again the most arduous computation in the saturated queue exponential approximation is the computation of b(s), the transform of the length of the busy period. As before, an approximation is proposed which uses a heavy traffic approximation to evaluate b(s) and b(0). Specifically, since $\nu = \lambda E[S] - 1 > 0$ let

$$\alpha(0) = \frac{-2\nu}{\sigma^2} \tag{43}$$

and

$$\alpha(s) = \frac{-\nu - \sqrt{\nu^2 + 2\sigma^2 s}}{\sigma^2} \tag{44}$$

and put

$$b(0) = E[e^{\alpha(0)S}] \tag{45}$$

and

$$b(s) = E[e^{\alpha(s)S}] \quad . \tag{46}$$

The remainder of the approximation is as in the exponential approximation for the saturated queue.

Table 5 shows results for both the "exact" exponential approximation and heavy traffic approximation for a saturated M/G/1 queue with $\lambda = 1$ and gamma service times having shape parameter 0.2 and scale parameter 7.5 and E[S] = 1.5. The true values are from Middleton (1979). As is expected the exponential approximation gives values which are closer to the true. Both approximations improve as t increases. The exponential approximation is very good for all times except the very smallest.

Table 5

Saturated M/G/1 Queue
$\lambda = 1$, Gamma Service Times, $E[S] = 1.5$
Shape parameter 0.2, Scale parameter 7.5

Time t	True $E[W_t]$	Exponential Approximation	Heavy Traffic Approximation
1.6	3.9	1.9	2.1
8.1	7.5	7.4	8.2
16.2	12.9	13.1	14.4
27.0	19.3	19.7	21.6
43.2	28.3	28.8	31.3
54.0	34.0	34.6	37.3
108.0	61.9	62.2	65.6
162.0	89.1	89.2	92.8
216.0	116.2	116.2	119.8
324.0	170.2	170.2	173.8
432.0	224.2	224.2	227.8

8. INFERENCE

The approximations to $E[W_t]$ can be used for inferential purposes by replacing moment and transforms by sample moments and empirical transforms. More specifically, suppose $\lambda = 1$ is known and service time data $d_1, \ldots, d_n$ are collected. The empirical Laplace transform of the service time distribution is

$$\phi_S(s) = \sum_{i=1}^{n} e^{-sd_i} \; ; \tag{47}$$

W_∞ is estimated by

$$\hat{W}_\infty = \frac{\lambda \overline{d^2}}{2(1 - \hat{\rho})} \tag{48}$$

where

$$\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i \; ; \tag{49}$$

$$\overline{d^2} = \frac{1}{n} \sum_{i=1}^{n} d_i^2 \; ; \tag{50}$$

and

$$\hat{\rho} = \lambda \bar{d} \; . \tag{51}$$

The empirical Laplace transform of the service times and the moment estimator of ρ, $\hat{\rho}$, can be used in formulas (2)-(4) to obtain an estimate of $\psi_W(s)$. This estimate, $\hat{\psi}_W(s)$, together with $\hat{W}$ can be used in the approximations to obtain estimates of $E[W(t)]$.

Tables 6-8 report results of simulation experiments to study the behavior of the estimates. Each simulation has 1000 replications. Each replication supposes a sample of 100 service times. The random numbers were generated using LLRANDOMII random number package, see Lewis and Uribe

Table 6

Moments of Estimates of $E[W_t]$
for
M/M/1 Queue with
$\lambda = 1 \quad E[S] = 0.9$

Time	True	Exponential		Heavy Traffic	
t	$E[W_t]$	Approximation		Approximation	
		Mean	Variance	Mean	Variance
6.5	1.9	1.8	0.1	2.0	0.1
32.4	4.2	4.3	2.0	4.5	2.4
97.2	5.9	6.8	14.0	7.1	15.6
162.0	6.7	8.3	33.1	8.6	35.9
486.0	7.9	13.0	210.6	13.2	218.5
648.0	8.0	14.7	347.6	14.9	357.5

Table 7

Moments of Estimates of $E[W_t]$
M/G/1 Queue
$\lambda = 1 \quad E[S] = 0.5$
Gamma Service Times
Shape = 0.2 Scale = 2.5

Time	True	Exponential		Heavy Traffic	
t	$E[W_t]$	Approximation		Approximation	
		Mean	Variance	Mean	Variance
0.3	0.13	0.12	0.001	0.13	.001
1.8	0.53	0.51	0.024	0.56	.03
3.0	0.71	0.70	0.058	0.78	.08
4.8	0.90	0.91	0.13	1.0	.17
9.0	1.2	1.2	0.33	1.3	.44
24.0	1.4	1.6	1.0	1.6	1.2

Table 8

Moments of Estimates of $E[W_t]$
M/G/1 Queue
$\lambda = 1$, $E[S] = 1.5$
Gamma Service Times
Shape = 0.2 Scale = 7.5

Time t	True $E[W_t]$	Exponential Approximation		Heavy Traffic Approximation	
		Mean	Variance	Mean	Variance
1.6	3.9	1.9	0.30	2.0	.31
8.1	7.5	7.3	6.7	8.1	7.6
54.0	34.0	34.5	256.8	36.8	277.7
162.0	89.1	90.1	2436.0	92.9	2488.0
324.0	170.2	171.3	10,210	174.3	10,296

(1981). If the true average service time is close to 1, then it is clearly possible for the sample traffic intensity $\hat{\rho}$ to be less than or greater than 1. If $\hat{\rho}$ is less than 1, then the estimate using the approximation for the stable queue is computed. If $\hat{\rho} > 1$, then the estimate using the approximation for the unstable queue is computed. This choice appears to be natural unless other information is available, or more assumptions are made.

In Tables 6-8 are reported means and variances of the estimates of $E[W_t]$ for the estimates based on the exponential approximation and the heavy traffic approximation.

In Table 6 results for a M/M/1 queue with $E[S] = .90$ are given. The same random numbers were used to compute the estimates for each of the times t. Of the 1000 replications, 137 had $\hat{\rho} > 1$ and so the unstable queueing approximations were used in these cases. Doubtless it is the contribution of these cases that lead to the pronounced over estimate of the mean.

In Table 7 are reported results for an M/G/1 queue with $\lambda = 1$ and gamma service times with $E[S] = .5$ having shape parameter .2 and scale parameter 2.5. None of the 1000 replications has $\hat{\rho} > 1$.

In Table 8 are reported moments for the estimates of $E[W_t]$ for an unstable M/G/1 queue with $\lambda = 1$ and gamma service times with shape parameter 0.2 and scale parameter 7.5 and $E[S] = 1.5$. Of the 1000 replications $\hat{\rho} > 1$ for 938 of them. The true values of $E[W_t]$ in all the tables are from Middleton (1979).

The means of the heavy traffic approximation are larger than those for the exponential approximation. The variances increase as t increases. The very large variances in Tables 6 and 8 are attributable to those estimates of $E[W_t]$ for which $\hat{\rho} > 1$.

In order to put the large variances of Table 8 into perspective, consider the following approximation.

As a first approximation, if $\lambda E[S] > 1$, then $E[W_t] \approx (\lambda E[S] - 1)t$ which can be estimated using the mean of the service times

$$\hat{m}(t) = (\lambda \bar{d} - 1)t \ .$$

The variance of $\hat{m}(t)$ for a sample size of 100 is

$$\mathrm{Var}[\hat{m}(t)] = \lambda^2 \frac{\mathrm{Var}(S)}{100} t^2 \ .$$

In the case of a gamma service time distribution with shape parameter 0.2 and scale parameter 7.5, $\lambda = 1$, $t = 162$

$$\mathrm{Var}[\hat{m}(t)] = 2952.5 \ .$$

Comparing this number to the corresponding variance of the estimate in Table 8 of 2436 indicates that the latter variance is not unreasonable.

The means of the estimates in Table 7 are close to the theoretical values in this stable queue. The means of the estimates in Table 6 are close to the theoretical values for the smaller t's. For larger t the means are greater than the theoretical values owing to those replications for which $\hat{\rho} > 1$.

The means of the estimates in the unstable queue case of Table 8 are close to the true values of $E[W_t]$ except for the smallest time, $t = 1.6$.

9. CONCLUSIONS

This paper proposes an easily-computed approximation to the finite-time expected waiting time for an M/G/1 system starting from an empty condition. Both unsaturated ($\rho < 1$) and saturated ($\rho > 1$) conditions are considered. Numerical evidence is presented to indicate that the quality of the approximation is usefully good, especially when ease of computation is an issue. Further, the methodology is adapted to assess expected waiting time when inferences must be made from a random sample of service times, and the decision is made to do so nonparametrically, i.e. without fitting a specific function. The results appear reasonable and potentially useful, and are not burdensome to obtain. The methodology investigated can also be applieɑ to the variety of queueing models that are close siblings of M/G/1: priority and breakdowns and "vacations" being examples. Of course other approximating and inferential options remain to be investigated.

References

J. Abate and W. Whitt. Transient behavior of the M/M/1 queue via Laplace transforms, Adv. Appl. Prob. 20 (1988), pp. 145-178.

J. Abate and W. Whitt. Approximations for the M/M/1 busy-period distribution. Unpublished paper, 1987.

S. Asmussen and H. Thorisson. Large deviation results for time-dependent queue length distributions. Commun. Statist.--Stochastic Models 4 (1988), pp. 99-116.

B. Efron. Bootstrap methods: another look at the jackknife. Annals of Statistics, Vol. 7, pp. 1-26.

W. Feller. An Introduction to Probability Theory and its Applications. Vol. II. John Wiley and Sons, Inc., New York, 1966.

D.P. Gaver and P.A. Jacobs. On inference and transient response for M/G/1 models. Teletraffic Analysis and Computer Performance Evaluation (Ed. O.J. Boxma, J.W. Cohen, and H.C. Tijms), Elsevier Science Publishers B.V. (North-Holland), 1986.

D.P. Gaver and P.A. Jacobs. System availability: time dependence and statistical inference by (semi) non-parametric methods. Naval Postgraduate School Technical Report, 1988.

D.P. Gaver. Observing stochastic processes and approximate transform inversion. Operations Research 15 (1966), pp. 444-459.

J. Keilson. Markov Chain Models--Rarity and Exponentiality, Springer-Verlag, New York, 1979.

P.A.W. Lewis and L. Uribe. The new Naval Postgraduate School random number package--LLRANDOM II. Naval Postgraduate School Technical Report NPS55-81-005, Monterey, 1981.

M.R. Middleton. Transient effects in M/G/1 queues: an empirical investigation. Technical Report No. 85, Department of Operations Research, Stanford University, Stanford, Calif. 1979.

A.R. Odoni and E. Roth. An empirical investigation of the transient behavior of stationary queueing systems. Operations Research 31 (1983), pp. 432-455.

H. Stehfest, Algorithm 368. Numerical inversion of Laplace transforms. Comm. Assoc. Comput. Mach. 13 (1970), pp. 47-49 [erratum 13, 624].

L. Takács. Introduction to the Theory of Queues. Oxford University Press, New York, 1962.

G.F. Newell. Applications of Queueing Theory. Chapman and Hall, London, 1982 (Second Edition).

DISTRIBUTION LIST

	No. of Copies
Library (Code 0142) Naval Postgraduate School Monterey, CA 93943-5000	2
Defense Technical Information Center Cameron Station Alexandria, VA 22314	2
Office of Research Administration (Code 012) Naval Postgraduate School Monterey, CA 93943-5000	1
Center for Naval Analyses 4401 Ford Avenue Alexandria, VA 22302-C268	1
Library (Code 55) Naval Postgraduate School Monterey, CA 93943-5000	1
Operations Research Center, Rm. E40-164 Massachusetts Institute of Technology Attn: R. C. Larson and J. F. Shapiro Cambridge, MA 02139	1
Koh Peng Kong OA Branch, DSO Ministry of Defense Blk 29 Middlesex Road SINGAPORE 1024	1
Arthur P. Hurter, Jr. Professor and Chairman Dept of Industrial Engineering and Management Sciences Northwestern University Evanston, IL 60201-9990	1
Institute for Defense Analysis 1800 North Beauregard Alexandria, VA 22311	1
Professor H. G. Daellenbach Department of Operations Research University of Canterbury Christchurch, NEW ZEALAND	1
Donald P. Gaver (Code 55Gv) Naval Postgraduate School Monterey, CA 93943-5000	10
Patricia A. Jacobs (Code 55Jc) Naval Postgraduate School Monterey, CA 93943-5000	10